Measuring the Microbiome: perspective on advances in exploring microbial life

James Foster[1], John Bunge[2], Jack A Gilbert[3,4], Jason Moore[5]

[1]Department of Biological Sciences, Institute for Bioinformatics and Evolutionary Studies, University of Idaho, Moscow, ID 83844-3051, U.S.A., [2]Department of Statistical Science, Cornell University, Ithaca, NY 14853, U.S.A., [3]Argonne National Laboratory, 9700 South Cass Avenue, Argonne, IL 60439, U.S.A. [4]Department of Ecology and Evolution, University of Chicago, 5640 South Ellis Avenue, Chicago, IL 60637, U.S.A. [5]Departments of Genetics and Community and Family Medicine, Institute for Quantitative Biomedical Sciences, Dartmouth College, Lebanon, NH 03756

Abstract

This article reviews recent advances in *microbiome studies*: molecular, statistical, and graphical {techniques} to gather data on, model, and explore how microbial organisms affect our environments and ourselves. Microbiome studies are in transition from surveys of which microbes are present in natural environments to quantifications of their community diversity and models of their ecological functions. We review the last 24 months of progress in this sort of research, and anticipate where the next two years will take us.

Introduction

We live in a microbial world, with microscopic organisms filling discrete ecosystems in such environments as soil, lakes and oceans, the human gut or skin, and even computer keyboards. Though microbiota include eubacteria, archea, viruses, and microscopic eukaria, we will consider only bacterial examples in this paper. Bacteria comprise most of the Earth's biomass and richness (1). They dominate such ecological factors as carbon cycling, greenhouse gas emission, and oxygen production. 90% of the cells in a human body are bacterial, as are 99% of the gene transcripts (2). However, most of the microbial world has been inaccessible to us, a kind of biological "dark matter," since we do not know how to culture over 97% of all bacteria, and since older cultivation independent microbial survey techniques such as TRFLP (Terminal Restriction Fragment Length Polymorphism), ARISA (Automated Intergenic Spacer Analysis), and gradient gel electophoresis have significant limitations. "Next Generation Sequencing" technologies have enabled, for the first time, high-throughput microbial sampling (3). The ultimate objective of microbiome studies is to understand the richness, structure, and function of microbial communities in nature.

Current microbiome studies begin by extracting DNA from a sample of the microbiome under study, post-processing the sequences to measure how many representatives of distinct populations (species, ecological functions, or other properties of interest) were observed in the sample, and then estimating a model of the original community. It is problematic to define a "microbiome", though for working purposes we take it to be a well-defined patch of an ecosystem, such as all bacteria in a prescribed sector of the ocean, all bacteria from a specific body part of all humans, or all viral particles in the gut of a particular pig. We use microbial ecology conventions rather than statistical ones, so that a "population" is a collection of all organisms of a given species, a "community" is a collection of "populations" that share a specific ecosystem, and a "sample" is a physical extract from a given microbiome.

Ambitious projects are underway to catalogue microbial life for the entire Earth (4), the ocean (5), and the human body (6). Surveys of transcriptomes and entire genomes have revealed more than half of all known protein sequences. Existing methods for estimating richness and community structure from observed samples are becoming more refined, improving model estimation, confidence quantification, and comparative methods (7-9). Finally, interactive, visual techniques are emerging with which to explore these complicated datasets prior to formal analysis.

Different sequencing technologies have idiosyncratic strengths and weaknesses, beyond the scope of this review, which are not fully understood (10). Currently most metagenetic researchers use the Roche 454 sequencing platforms, and metagenomic/metratranscriptomic researchers also make extensive use of the Illumina GAIIx/HiSeq2000. The Roche 454 GS-FLX Titanium can now generate in excess of 1 million reads per run, which takes 23 hours, with read lengths up to 1000 bp (average around 500bp); the average run generates 750 Mbp of sequencing

data. The Illumina HiSeq2000 platform can now generate approximately 4 billion reads per run, which takes 10 days, with (usually) 150bp paired-end reads to create an approximately 250bp product; the average run generates 1 Tbp of sequencing data. The vendors reported these data at the time of this writing. Newer technologies, such as single molecule sequencing and smaller single lab devices are not widely used yet, and Sanger sequencing of large-insert libraries are still significant (11).

Recent bioinformatics advances have significantly improved sequencing and assembly errors detection and correction. Several packages provide pipelines to bring these new algorithms into the lab (12,13). Bioinformaticists continue to improve algorithms for detecting specific types of error, such as chimeric sequences (14) and precise but inaccurate reads (15,16).

This review surveys recent advances in efforts to measure the diversity of complete microbial communities. We limit references to recent publications that serve as jumping off points for further exploration, rather than a complete literature survey. First we discuss studies based on metagenetic data, from amplicons of single genes or genetic regions. Next, we review analyses of metagenomic and metatranscriptomic data, from shotgun sequencing of multiple genomes or genome transcripts. We then review advances and limitations in statistical techniques for diversity estimation. Then we discuss visual analytics, hypothesis generation by visually exploring these very large sequence datasets. Finally, we speculate on how microbiome studies may change in the next two years.

Metagenetics, amplicon sequencing

Hypervariable regions of individual, highly conserved genes, such as the small ribosomal subunit in non-eukaryotes, have served as proxies for species since Woese and Fox first used them to demonstrate that archea were a separate kingdom (17-19). With new sequencing

technologies it became possible to sample all the 16S genes in a specimen without having to isolate and cultivate organisms in order to amplify DNA separately. By tagging specimens with molecular barcodes, labs can multiplex several treatments and controls into a single sequencing run, making it possible to survey and compare different microbiomes with very few sequencing jobs, dramatically shrinking the time between sample preparation and data analysis and the sequencing costs.

The 16S rRNA gene remains a good, but far from ideal, molecular marker for microbial diversity. Hundreds of thousands of 16S rRNA genes have been fully sequenced and classified (13,20). As with all databases, ribosomal databases are growing larger and better, so analysis relying on them can only improve. The secondary structure of the 16S rRNA molecule is well characterized, at least for reference strains, which makes is possible to perform fast, secondary structure driven alignments (21,22). However, the diversity of the 16S gene does not always reflect phylogenetic relationships or metabolic potential that are known from other sources (23). Existing databases rarely classify below the family level, with results often reported as the order or even phyla level, even though different species or even strains are likely to have very different roles in microbiomes. Database sequences are surely biased samples of reality, since they assume at least that their targets are amenable to existing sequencing and annotation methodologies. They have been further biased by our fixation on potential pathogens and environmental contaminants. However, the 16S gene is likely to remain the most reliable and broadly applicable marker for some time.

To date, sequencing technologies have limited 16S sampling to small fragments, rather than entire genes or genomes. Primers exist for hypervariable regions known as V1 through V9, of widely varying lengths and phylogenetic resolution (24). Different regions, and combinations

of regions, have different strengths and weaknesses (25-27). Historically, human microbiome surveys typically sample from regions near V3, while environmental surveys often sample from regions near V6. As sequencing technologies and protocols improve, projects are sequencing longer regions, such as V3-V5 (from the beginning of V3 to the end of V5) or V6-V9 (from V6 to V9). Eventually, it may become routine to use the entire 16S gene, multiple marker genes, or even entire genomes.

There are two types of algorithms for inferring microbiome diversity and structure from "clean" sequences, and both have improved greatly in the last two years.

Clustering methods group sequences by similarity, computing statistics from the number and size of clusters. Clustering methods are biased by how one measures similarity and what similarity threshold one uses (25,28). Older distance clustering methods begin by computing similarities for all pairs of sequences, producing massive distance matrices. Newer algorithms compute clusters on the fly, requiring far less computer memory. Clusters are often called Operational Taxonomic Units (OTUs), a term borrowed from systematics, though they represent sequence similarity, which may no reflect organismal phylogeny or functional diversity. Recent studies have shown that, in general, average neighbor clustering (usually at a 97% similarity threshold) following single linkage clustering (usually at a 98% similarity threshold) works better for estimating community diversity than alternatives (16). UNIFRAC algorithms estimate phylogenetic divergence as a similarity measure, providing the best current method for estimating between-population (so called beta) diversity(29). Very few algorithms exist that rigorously fit statistical models to sequence data in order to estimate microbiome structure (see below).

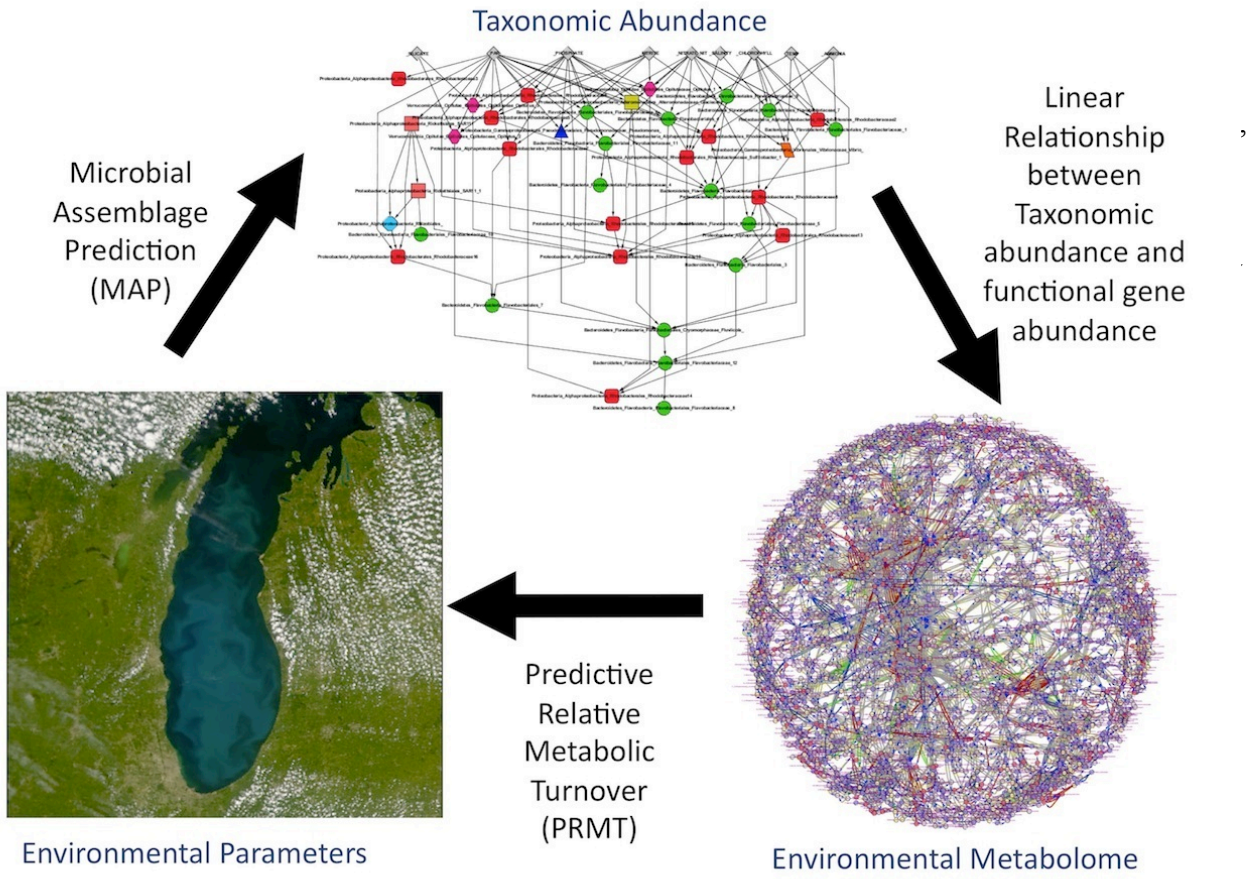

Figure 1. Microbial Assemblage Prediction (MAP - Larsen & Gilbert, Nature Methods, Submitted) and Predicted Relative Metabolite Turnover (PRMT - Larsen, Microbial Informatics and Experimentation, 2011) use environmental parameters to predict how microbial community structure will change in response to relative turnover of over 1000 metabolites in the ecosystem.

Researchers use metagenomic and metatranscriptomic sequencing to explore the functional and expressed potentials of microbial communities. Most studies have performed extensive sequencing of bacterial communities (33). But viral (34) and eukaryotic (35) communities have also been studied. Indeed, recent metagenomic data analysis may have discovered a fourth domain of life (36).

The difficulty of assembling and annotating the data, due to short read lengths, has been the primary challenge to analyzing high throughput metagenomic/metatranscriptomic data (37). Assembly is important for the reconstruction of genes and operons for functional assignation and improved annotation of taxonomy (38), but also for re-assembly of whole genomes from metagenomic DNA (39). Independently of assembly problems, functional annotation of function is a difficult problem, compounded by the sheer quantity of sequence data. Consequently, automated annotation has become routine, with little or no manual assessment of accuracy (40). One of the most appropriate ways of defining the accuracy of assembly and annotation of metagenomic data is to use in silico simulated data from fragmented genomes (41) or actual fragmented genomic DNA from known organisms (42).

Nonetheless, comparative metagenomics remains one of the most powerful ways to explore gene distribution across different ecosystems (43). Several tools and technologies exist for comparing functional community dynamics across different metagenomic datasets (44). Current techniques are limited by difficulties contextualizing sequencing data with environmental metadata from the target ecosystem (45). However, techniques are being developed to improve analyses, once environmental metadata about the niche space in which the community was structured is available (46).

It is possible to model complex community dynamics in relation to the chemical and physical dynamics of the ecosystem (Figure 1), even without exhaustive sequence and environmental data. For example, tools exist to derive the abundance of gene/transcript fragments annotated to known enzyme activities from metagenomic and metatranscriptomic data (46). Tools are under development to predict how bacterial community structure will respond to relative changes in the consumption or production of these metabolites, given metagenetic data for the community (Microbial Assemblage Prediction (MAP) algorithm: Larsen & Gilbert, Nature Methods, In submission).

Statistics for diversity estimation

The statistical challenges for microbiome studies are to estimate population richness and diversity, model community structure, quantify uncertainty, and compare estimates rigorously (47). Most current techniques begin with frequency count data, which groups observations into bins and report the number of members of each bin.

There are two main approaches to richness estimation from count data. The *classical* or *frequentist* approach is better represented in both the literature and in available software. Coverage-based nonparametric estimators like Chao and ACE are popular, being simple to compute, and are available in most metagenetic bioinformatics packages such as Mothur and QIIME (12,48). But they are known to be biased downward in high-diversity situations, and to behave erratically when outliers are present (47). Recently more stable but computationally intensive parametric mixture models have been introduced. Both types of estimate are available in a single package, CatchAll (7). Further, CatchAll computes several different estimates and returns a ranked comparison of the "best" analyses for a given dataset.

The *Bayesian* approach, by contrast, begins with a prior probability distribution that represents what is known or believed about the diversity before collecting any data. Using Bayes' Theorem, this approach then derives a posterior distribution using the observed data, which yields the final estimate of diversity along with error terms and confidence intervals. There are two ways to define the prior. In *objective* or *noninformative* Bayesian analysis one minimizes the amount of information in the prior so that it influences the end result as little as possible; while in *subjective* or *informative* Bayesian analysis the prior expresses the experimenter's beliefs about the diversity, or weights the results according to known factors that are unrelated to the observed data. Both have been studied in the diversity estimation literature, but the objective Bayesian approach is more widely accepted (49,50). Indeed it promises to be statistically and computationally stable and flexible, and may well be a strong competitor to the frequentist methods. But at present there is no simple and generally accessible Bayesian diversity estimation software, so we have less applied experience than for the classical approach.

Recently, statistical methods have been developed that adjust estimates according to patterns in or assumptions about the frequency count data. For example, the successive ratios of frequencies (the number of doubletons divided by the number of singletons, tripletons divided by doubletons, etc.) have known statistical properties, which led to a new estimation method (available in CatchAll) (51)**.** Another example incorporates the suspected unreliability of low frequency counts into diversity estimates. Recent analyses of artificially constructed communities with known diversity and structure indicate that existing methods may systematically lead to inflated low frequency counts. Strategies to address such biases include: (i) using a Bayesian prior weighted toward lower diversity values; (ii) reporting lower bounds rather than direct estimates for the total diversity; (iii) statistically separating the projected population

into low and high-diversity components and deleting or downweighting the latter; and (iv) by pooling low frequency counts up to some cutoff (say, the singletons and doubletons) and re-estimating the total diversity from these left-censored data (52). All of these strategies are statistically feasible, although not all have been implemented in software (CatchAll includes (ii) and (iii)), and this remains an area of current research.

The next logical step is to move from estimating the diversity of a single population to comparing diversity levels across two or more populations. Given reliable richness estimates for individual communities, it is straightforward to make statistical comparisons of richness between microbiomes. It is considerably more challenging to quantify how much population structure is shared between two or more communities. One common metric for two communities is the Jaccard index, which is the ratio of the number of shared populations to the total number of populations observed. Other between population diversity metrics include Sørensen, Bray-Curtis, and Morisita-Horn (48)**.** However, these formulae are often applied to compare observed samples rather than estimated communities, leading to statistically indefensible practices such as discarding data to "normalize" samples to the same size. What is lacking is between-population diversity metrics that account for both observed and unobserved populations. This appears to be a difficult statistical problem. Chao et al. provided a nonparametric estimator of the (population) Jaccard and Sørensen indices (8) but few other solutions have been proposed (53).

Finally, microbiome studies need to model or predict richness and diversity using covariate data, such as observable biological, chemical, or other environmental variables. If the response or dependent variable is simply the (estimated) richness then standard statistical modeling techniques such as regression are appropriate. But modeling diversity and structure,

rather than mere not just richness, as a function of the predictors, requires techniques such as canonical correspondence analysis (9).

Moreover, such analyses should be based on estimates of unobserved structure, rather than exclusively observed data, since substantial unobserved diversity is typical of microbial ecology studies.

Visualizing the results

Microbiome data are inherently high dimensional and complex. Suppose the goal of a project is to relate bacteria community structure at a particular body site to clinical observations. A typical data set might include a list of hundreds of bacterial species that are hierarchically organized into different groups, including genus, families, orders classes and phyla. This is further complicated with information about genes and pathways that are present in each of the bacterial species and with how these relate to clinical endpoints. The genomic information of the host, such as demographic data, patient specifics, and lifestyle data may also be important. The ultimate challenge is to put these many different layers of information together in a statistical or machine learning analysis to identify clinically useful patterns.

Given this level of data complexity, it is important for the researcher to have tools with which to visualize and explore data. Visual interaction allows the researcher to critically explore the measurements themselves for quality control, for discovering patterns that lead to new hypotheses, and for interpreting results. Also, it is often desirable to communicate results visually to other scientists and clinicians. However, it is challenging to choose the right visualization technique for the right type of data or information, given that there are so many information visualization methods (54).

The close integration of computational analysis with visualization and human-computer interaction is an emerging discipline called *visual analytics* (55). This is distinguished from scientific visualization that focuses on the mathematics and physics of visualizing 3D objects, and information visualization that focuses on methods such as heat maps for showing high-dimensional research results. What distinguishes visual analytics is the integration of data analysis with visualization methods such that data analysis can be launched directly from the visualization and the visualization adjusted in response to the data analysis. Computer hardware technology that makes it easy for the user to interact with the software, such as the Microsoft Surface Computer or the Apple iPad, enable visual analytics. All of this combined with a 3D visualization screen or wall provides a modern visual analytics discovery environment that immerses the user in their data and research results.

Several different information visualization methods have been useful for the analysis of microbiome data. For example, heat maps, introduced more than fifty years ago (56), have become a popular and useful for visualizing population structure in large microbial communities and for clusters of expression patters in genomics (57). A heat map consists of a 2D grid or matrix of colored squares where each square represents an observation of a random variable and the color of the square is proportional to the value of that observation. It is common to order the squares along the two axes with additional categorical data such as bacterial phyla and tissue type. For example, a recent study by Wu et al. explored the relationship between long-term dietary patterns and gut microbial enterotypes (58). This study used Spearman correlations to estimate the association between different nutrients and bacterial genera in 98 healthy volunteers. It summarized the results with a heat map, where each column represented different taxa, each row represented a different nutrient and the color of each square represented the magnitude of

the correlation, with darker red representing stronger positive correlations and darker blue representing stronger negative correlations. The authors also performed a hierarchical cluster analysis to organize the results into visual patterns that were easier to interpret. For example, the authors found that fat-related nutrients tended to be more similar in the correlations across taxa than other nutrient groups. In addition to heat maps, the authors also used principal components analysis (PCA) to identify linear combinations of gut microbial taxa associated with long-term diet. They used 2D and 3D scatterplots to identify clusters of patients defined by the first two or three principal components. This type of multivariate analysis is inherently visual and can prove to be a very useful information visualization tool for microbiome analysis.

Visual analytics, combining information visualization with human-computer interaction, is emerging as a useful tool for microbiome visualization and analysis. For example, Moore et al. integrate and visualize several additional dimensions of information by adding a z-axis to a traditional heat map (59). In this study, the authors implemented the 3D heat map using a commercial 3D video game engine called Unity 3D. The video game engine makes it possible to interact with the 3D visualization as you would in a video game. This open-source software combines human-computer interaction and visualization in a 3D Heat Map in a way that is not possible with common analysis tools such as Microsoft Excel or R. The second example comes from a human microbiome study of (60). Here, the authors use movies to visualize the temporal variation in the vaginal microbiome of 396 women from different racial groups, and work is underway to incorporate temporal and patient metadata. The use of movies allows users to interact with the visualization in a way that is not possible with static images.

Where things are going

Sequencing technologies will continue to improve in both accuracy and throughput, and benchtop sequencers will become standard equipment in individual labs. Metagenetics will rely more on whole gene samples, removing the bias associated with selection a gene fragment. In particular, complete 16S DNA sequences will become standard for microbial systematics. However, we anticipate that metagenetics will become a quick screening technique, preliminary to more detailed metagenomic studies.

The ideal dataset for genomic-based microbial studies of any given ecosystem, including those associated with animals, such as humans, is a complete genome for every single organisms at a given time in the ecosystem. These data would make it possible to completely characterize the genetic diversity of the system, to generate comprehensive models of microbial metabolism and interactions, and to design experiments that manipulate the system by adding or removing specific populations. The most obvious route towards this is single genome isolation and sequencing, which is currently performed by isolating single microbial cells and sequencing them directly (61) to identify the functional potential of the organisms, and to design economically feasible, rather than exhaustive, shotgun metagenomic.

But the ultimate objective is to build complete, predictive models of how microbiomes interact and how they respond to stimuli such as climate change, agricultural practices, and disease(62). Parameterizing such complex models will continue to require metatranscriptomic and other “omic” studies of the expressed capability of community members (33,63,64). Using techniques such as autonomous collection and preservation of microbial communities for metatranscriptomic analysis combined with quantitative characterization of transcription in

metatranscriptomic data we may start to see a revolution in our ability to resolve functional capability(65,66).

Statistical techniques are likely for improved parametric model estimation, error and uncertainty bounds, and comparing diversity statistics. These are likely to include refined techniques for censoring unreliable data, without first characterizing where the noise comes from. We also anticipate that software tools will become more available for sophisticated analyses, but that interpreting results will still require statistical expertise.

Information visualization and visual analytics will become standard parts of microbiome research workflows. Integration into statistical computing software such as R is already underway, so that analyses can be launched directly from the visualization applications. The ability to launch statistical analyses directly from the visualization environment opens the door to making discoveries that are inspired by visual cues rather than pre-conceived hypotheses that are dependent on existing knowledge. Perer and Shneiderman (2009) present design guidelines for evaluating visual analytics software(59).

Acknowledgements

This work was funded in part by: NIH COBRE grant P20RR16448, NIH INBRE grant P20RR016454, and NSF STC "BEACON Center for the Study of Evolution in Action" NSF STC DBI-0939454 to JAF; National Science Foundation grant DEB-08-16638 to JB; U.S. Dept. of Energy Contract DE-AC02-06CH11357 to JAG; and NIH grants LM009012, LM010098 and AI59694 to JHM. This research was conducted using the resources of the Cornell Center for Advanced Computing, which receives funding from Cornell University, the National Science Foundation, and other leading public agencies, foundations, and corporations. We also thank the

organizers of the Pacific Symposium on Biocomputing for their special session on "Microbiome Studies".

Annotated references

1. Whitman WB, Coleman DC, Wiebe WJ. Prokaryotes: the unseen majority. Proc Natl Acad Sci U S A. 1998 Jun. 9;95(12):6578–6583.

2. Sears CL. A dynamic partnership: celebrating our gut flora. Anaerobe. 2005 Oct.;11(5):247–251.

3. Rodrigue S, Materna AC, Timberlake SC, Blackburn MC, Malmstrom RR, Alm EJ, et al. Unlocking short read sequencing for metagenomics. PLoS ONE. 2010;5(7):e11840.

4. Gilbert JA, Meyer F, Jansson J, Gordon J, Pace N, Tiedje J, et al. The Earth Microbiome Project: Meeting report of the "1 EMP meeting on sample selection and acquisition" at Argonne National Laboratory October 6 2010. Stand Genomic Sci. 2010;3(3):249–253.

The issue of ecosystem characterization is well demonstrated in this manuscript, which leverages metagenomics, metatranscriptomics and amplicon taxonomic analysis for multiple domains. The study identifies ecosystem dynamics over diel and seasonal time scales for both taxa, genes and transcripts, and helps to shape our understanding of the reaction of microbes to change.

5. Bowler C, Karl D. Microbial oceanography in a sea of opportunity. Nature. 2009.

6. Brüls T, Weissenbach J. The human metagenome: our other genome? Hum Mol Genet. 2011 Aug. 31.

7. Bunge J. Estimating the number of species with CatchAll. Pac Symp Biocomput. 2011.

Original introduction to software CatchAll, including operational description and references to theoretical underpinnings for all methods.

8. Chao A, Chazdon RL, Colwell RK, Shen T-J. Abundance-based similarity indices and their estimation when there are unseen species in samples. Biometrics. 2006 Jun.;62(2):361–371.

9. Legendre P, Legendre L. Numerical Ecology, Volume 24, Third Edition (Developments in Environmental Modelling). 3rd ed. Elsevier; 2012.

The classic reference on methods for quantifying ecological diversity.

10. Suzuki S, Ono N, Furusawa C, Ying B-W, Yomo T. Comparison of sequence reads obtained from three next-generation sequencing platforms. PLoS ONE. 2011;6(5):e19534.

11. Brochier-Armanet C, Deschamps P, López-García P, Zivanovic Y, Rodriguez-Valera F, Moreira D. Complete-fosmid and fosmid-end sequences reveal frequent horizontal gene transfers in marine uncultured planktonic archaea. ISME J. 2011 Aug.;5(8):1291–1302.

12. Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, et al. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl Environ Microbiol. 2009 Dec.;75(23):7537–7541.

Mothur is a C++ package with current implementations of most microbiome bioinformatics tools. It is comprehensive, but has no GUI.

13. Cole JR, Chai B, Farris RJ, Wang Q, Kulam-Syed-Mohideen AS, McGarrell DM, et al. The ribosomal database project (RDP-II): introducing myRDP space and quality controlled public data. Nucleic Acids Res. 2007 Jan.;35(Database issue):D169–72.

The RDP offers several tools for microbiome analysis, in addition to the most current collection of rRNA gene sequences

14. Edgar RC, Haas BJ, Clemente JC, Quince C, Knight R. UCHIME improves sensitivity and speed of chimera detection. Bioinformatics. 2011 Aug. 15;27(16):2194–2200.

15. Reeder J, Knight R. Rapidly denoising pyrosequencing amplicon reads by exploiting rank-abundance distributions. Nat Methods. 2010 Sep. 1;7(9):668–669.

16. Huse SM, Welch DM, Morrison HG, Sogin ML. Ironing out the wrinkles in the rare biosphere through improved OTU clustering. Environ Microbiol. 2010 Jul.;12(7):1889–1898.

17. Woese CR, Fox GE. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proc Natl Acad Sci U S A. 1977 Nov.;74(11):5088–5090.

18. Woese CR. A new biology for a new century. Microbiol Mol Biol Rev. 2004 Jun. 1;68(2):173–186.

19. Schmidt TM. The maturing of microbial ecology. Int Microbiol. 2006 Aug. 31;9(3):217–223.

20. Pruesse E, Quast C, Knittel K, Fuchs BM, Ludwig W, Peplies J, et al. SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence

data compatible with ARB. Nucleic Acids Res. 2007;35(21):7188–7196.

21. Nawrocki E, Kolbe D, Eddy S. Infernal 1.0: inference of RNA alignments. Bioinformatics. 2009 Mar. 23.

22. Birin H, Gal-Or Z, Elias I, Tuller T. Inferring Horizontal Transfers in the Presence of Rearrangements by the Minimum Evolution Criterion. Bioinformatics. 2008 Feb. 4.

23. Kuczynski J, Liu Z, Lozupone C, McDonald D, Fierer N, Knight R. Microbial community resemblance methods differ in their ability to detect biologically relevant patterns. Nat Methods. 2010 Oct. 1;7(10):813–819.

24. Kim M, Morrison M, Yu Z. Evaluation of different partial 16S rRNA gene sequence regions for phylogenetic analysis of microbiomes. 2011;84(1):81–87.

25. Schloss P. The Effects of Alignment Quality, Distance Calculation Method, Sequence Filtering, and Region on the Analysis of 16S rRNA Gene-Based Studies. PLoS Comput Biol. 2010.

A thorough comparison of details in a standard pipeline for metagenetic analysis, based on Mothur.

26. Liu B, gordon T, Ghodsi M, Pop M. MetaPhyler: Taxonomic Profiling for Metagenomic Sequences. manuscript. 2010 Oct. 28;:1–6.

27. Wommack K, Bhavsar J, Ravel J. Metagenomics: Read length matters. Appl Environ Microbiol. 2008 Jan. 11.

28. Sun Y, Cai Y, Huse SM, Knight R, Farmerie WG, Wang X, et al. A large-scale benchmark study of existing algorithms for taxonomy-independent microbial community analysis. Briefings in Bioinformatics. 2011 Apr. 27.

29. Lozupone C, Lladser M, Knights D, Stombaugh J, Knight R. UniFrac: an effective distance metric for microbial community comparison. ISME J. 2010.

30. DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, et al. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl Environ Microbiol. 2006 Jul.;72(7):5069–5072.

31. Knights D, Costello E. Supervised classification of human microbiota. 2011. FEMS microbiology, 35(2):343–359.

32. Huse SM, Huber JA, Morrison HG, Sogin ML, Welch DM. Accuracy and quality of massively parallel DNA pyrosequencing. Genome Biol. 2007;8(7):R143.

33. Gilbert JA, Field D, Swift P, Thomas S, Cummings D, Temperton B, et al. The taxonomic and functional diversity of microbes at a temperate coastal site: a “multi-omic” study of

seasonal and diel temporal variation. PLoS ONE. 2010;5(11):e15545.

The issue of ecosystem characterization is well demonstrated in this manuscript, which leverages metagenomics, metatranscriptomics and amplicon taxonomic analysis for multiple domains. The study identifies ecosystem dynamics over diel and seasonal time scales for both taxa, genes and transcripts, and helps to shape our understanding of the reaction of microbes to change.

34. Kristensen DM, Mushegian AR, Dolja VV, Koonin EV. New dimensions of the virus world discovered through metagenomics. Trends Microbiol. 2010 Jan.;18(1):11–19.

35. Cuvelier ML, Allen AE, Monier A, McCrow JP, Messié M, Tringe SG, et al. Targeted metagenomics and ecology of globally important uncultured eukaryotic phytoplankton. Proc Natl Acad Sci USA. 2010 Aug. 17;107(33):14679–14684.

36. Wu D, Wu M, Halpern A, Rusch DB, Yooseph S, Frazier M, et al. Stalking the Fourth Domain in Metagenomic Data: Searching for, Discovering, and Interpreting Novel, Deep Branches in Marker Gene Phylogenetic Trees. PLoS ONE. 2011 Mar. 18;6(3):e18011.

Translating metagenomic data into phylogenetic information is a key methodological development. In this paper, the authors use taxonomic-marker genes from un-biased metagenomic data to identify uncharacterized viruses and ancient paralogs not yet seen in any cultured organism.

37. Wooley JC, Godzik A, Friedberg I. A primer on metagenomics. PLoS Comput Biol. 2010 Feb.;6(2):e1000667.

38. Warren RL, Holt RA. Targeted assembly of short sequence reads. PLoS ONE. 2011;6(5):e19816.

39. Chitsaz H, Yee-Greenbaum JL, Tesler G, Lombardo M-J, Dupont CL, Badger JH, et al. Efficient de novo assembly of single-cell bacterial genomes from short-read data sets. Nat Biotechnol. 2011 Sep. 18.

Using single cell isolation, DNA amplification and short-read sequencing technologies to assembly genomes of microbes isolates from the environment is a holy grail of metagenomics. This paper highlights the state of the art in this technological capability.

40. Temperton B, Gilbert JA, Quinn JP, McGrath JW. Novel analysis of oceanic surface water metagenomes suggests importance of polyphosphate metabolism in oligotrophic environments. PLoS ONE. 2011;6(1):e16499.

41. Pignatelli M, Moya A. Evaluating the fidelity of de novo short read metagenomic assembly using simulated data. PLoS ONE. 2011;6(5):e19984.

42. Morgan JL, Darling AE, Eisen JA. Metagenomic sequencing of an in vitro-simulated microbial community. PLoS ONE. 2010;5(4):e10209.

43. Biddle J, White J, Teske A. Metagenomics of the subsurface Brazos-Trinity Basin (IODP site 1320): comparison with other sediment and pyrosequenced metagenomes. ISME J. 2011.

44. Suparna Mitra, Paul Rupek, … Daniel H Huson Functional analysis of metagenomes and metatranscriptomes using SEED and KEGG. BMC Bioinformatics. 2011;12 Suppl 1:S21.

45. Yilmaz P, Kottmann R, Field D, Knight R, Cole JR, Amaral-Zettler L, et al. Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. Nat Biotechnol. 2011 May;29(5):415–420.

46. Weber M, Teeling H, Huang S, Waldmann J. TaxSOM: Application of Self-Organizing Maps to link biodiversity and functional data in environmental metagenomics. ISME journal. 2011.

47. Bunge J. Statistical estimation of uncultivated microbial diversity. Uncultivated Microorganisms. 2009.

48. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, et al. QIIME allows analysis of high-throughput community sequencing data. Nat Methods. 2010 May;7(5):335–336.

QIIME implements the best current algorithms for most of the steps in a metagenetics pipeline. It is very fast and highly recommended.

49. Barger K. Objective Bayesian Estimation for the Number of Species. Bayesian Analysis. 2010.

Theoretical investigation of objective Bayes methods for diversity estimation, with applications in microbial ecology and other fields.

50. Quince C, Curtis TP, Sloan WT. The rational exploration of microbial diversity. ISME J. 2008 Oct.;2(10):997–1006.

Quasi-objective/noninformative Bayesian methodology applied to microbial richness estimation. Extensive exploration of method.

51. Rocchetti I, Bunge J. Population size estimation based upon ratios of recapture probabilities. arXiv. 2011 Jul. 27;stat.AP.

52. Bunge J, Böhning D, Allen H, Foster JA. Estimating population diversity with unreliable low frequency counts. In: B, editor. Pacific Symposium in Biocomputing. Kohala, HI: 2012. p. 1–10.

53. Engen S, Grøtan V. Estimating similarity of communities: a parametric approach to spatio−temporal analysis of species diversity. Ecography. 2010.

54. Heer J, Bostock M, Ogievetsky V. A tour through the visualization zoo. Communications of the ACM. 2010 Jun.;53(6).

55. Zhang Q, Segall R, Cao M. Visual Analytics and Interactive Technologies. Information Science Reference; 2010.

56. Sneath PH. The application of computers to taxonomy. J. Gen. Microbiol. 1957 Aug.;17(1):201–226.

57. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci U S A. 1998 Dec. 8;95(25):14863–14868.

58. Wu GD, Chen J, Hoffmann C, Bittinger K, Chen Y-Y, Keilbaugh SA, et al. Linking Long-Term Dietary Patterns with Gut Microbial Enterotypes. Science. 2011 Sep. 1.

59. Moore J and Lari R, Human microbiome visualization using 3D technology. Pacific Symposium on Biocomputing. In Press. 2011.

60. Ravel J, Gajer P, Abdo Z, Schneider GM, Koenig SSK, McCulle SL, et al. Vaginal microbiome of reproductive-age women. Proc Natl Acad Sci USA. 2011 Mar. 15;108 Suppl 1:4680–4687.

61. Woyke T, Tighe D, Mavromatis K, Clum A, Copeland A, Schackwitz W, et al. One bacterial cell, one complete genome. PLoS ONE. 2010;5(4):e10314.

62. Denef VJ, Mueller RS, Banfield JF. AMD biofilms: using model communities to study microbial evolution and ecological complexity in nature. ISME J. 2010 May;4(5):599–610.

63. Gilbert JA, Steele JA, Caporaso JG, Steinbrück L, Reeder J, Ben Temperton, et al. Defining seasonal marine microbial community dynamics. ISME J. 2011 Aug. 18;:1–11.

64. Gosalbes MJ, Durbán A, Pignatelli M, Abellan JJ, Jiménez-Hernández N, Pérez-Cobas AE, et al. Metatranscriptomic approach to analyze the functional human gut microbiota.

PLoS ONE. 2011;6(3):e17447.

65. Ottesen EA, Marin R, Preston CM, Young CR, Ryan JP, Scholin CA, et al. Metatranscriptomic analysis of autonomously collected and preserved marine bacterioplankton. ISME J. 2011 Jun. 30.

66. Field D, Amaral-Zettler L, Cochrane G, Cole JR, Dawyndt P, Garrity GM, et al. The Genomic Standards Consortium. PLoS Biol. 2011 Jun.;9(6):e1001088.

Keywords

Microbial ecology, biodiversity, metagenomics, next generation sequencing, microbiome, visual analytics

Author Biographies

Dr. Foster is a Professor in the Department of Biological Sciences and the Institute for Bioinformatics and Evolutionary STudies (IBEST) at the University of Idaho.

Dr. Bunge is Associate Professor in the Dept. of Statistical Science at Cornell University.

Dr Gilbert is an Environmental Microbiologist at Argonne National Laboratory, and Assistant Professor in the Department of Ecology and Evolution at University of Chicago.

Dr. Moore is the Third Century Professor of Genetics and Community and Family Medicine and Director of the Institute for Quantitative Biomedical Sciences at Dartmouth College.